Theoretical Validation of Potential Habitability via Analytical and Boosted Tree Methods: An Optimistic Study on Recently Discovered Exoplanets

Snehanshu Saha[1], Suryoday Basak[1], Margarita Safonova[3], Kakoli Bora[2],
Surbhi Agrawal[1], Poulami Sarkar[1] and Jayant Murthy[4]

Abstract— **Seven Earth-sized planets, known as the TRAPPIST-1 system, was discovered with great fanfare in the last week of February 2017. Three of these planets are in the habitable zone of their star, making them potentially habitable planets a mere 40 light years away. Discovery of the closest potentially habitable planet to us just a year before – Proxima b and a realization that Earth-type planets in circumstellar habitable zones are a common occurrence provides the impetus to the existing pursuit for life outside the Solar System. The search for life has two goals essentially: looking for planets with Earth-like conditions (Earth similarity) and looking for the possibility of life in some form (habitability). An index was recently proposed, Cobb-Douglas Habitability Score (CDHS), based on Cobb-Douglas habitability production function (CD-HPF), which computes the habitability score by using measured and estimated planetary parameters. As an initial set, radius, density, escape velocity and surface temperature of a planet was used. The proposed metric, with exponents accounting for metric elasticity, is endowed with analytical properties that ensure global optima and can be scaled to accommodate a finite number of input parameters. We show here that the model is elastic, and the conditions on elasticity to ensure global maxima can scale as the number of predictor parameters increase. K-NN (K-Nearest Neighbor) classification algorithm, embellished with probabilistic herding and thresholding restriction, utilizes CDHS scores and labels exoplanets to appropriate classes via feature-learning methods yielding granular clusters of habitability. The algorithm works on top of a decision-theoretical model using the power of convex optimization and machine learning. The goal is to classify the recently discovered exoplanets into the "Earth League" and several other classes. A second approach, based on a novel feature-learning and tree-building method classifies the same planets without computing the CDHS of the planets and produces a similar outcome. The convergence of the two different approaches indicates the strength of the proposed solution scheme and the likelihood of the potential habitability of the recently announced discoveries.**

I. Introduction

With discoveries of exoplanets pouring in hundreds, it is becoming necessary to develop some sort of a quick screening tool – a ranking scale – for evaluating habitability perspectives for the follow-up targets. We have proposed a novel inductive approach, inspired by the Cobb-Douglas model from production economics, to verify theoretical conditions of global optima of the functional form to model and to compute the habitability score of exoplanets – the Cobb-Douglas Habitability Score (CDHS; [1]). While our paper "CD-HPF: New Habitability Score Via Data Analytic Modeling" was in production, the discovery of an exoplanet Proxima b orbiting the nearest star (Proxima Centauri) to the Sun was announced [2]. This planet generated a lot of stir in the news [3] because it is located in the habitable zone and its mass is in the Earth's mass range: $1.27 - 3$ $\mathrm{M}_{\oplus}$, making it a potentially habitable planet (PHP) and an immediate destination for the Breakthrough Starshot initiative [4]. A few months after the announcement of Proxima b, another family of terrestrial-size exoplanets – the TRAPPIST-1 system – was discovered [5].

This work is motivated by testing the efficacy of the suggested model, CDHS, in determining the habitability score, the proximity to the "Earth-League", of the recently discovered Proxima b. The habitability score model has been found to work well in classifying previously known exoplanets in terms of potential habitability. Therefore it was natural to test whether the model can also classify it as potentially habitable by computing its habitability score. This could indicate whether the model may be extended for a quick check of the potential habitability of newly discovered exoplanets in general. As we see in Section VI, this is indeed the case with the TRAPPIST-1 planets.

The flowchart in Figure 1 summarizes our new approach to the habitability investigation of exoplanets. (on the example of Proxima b and TRAPPIST-1 system). This approach is based on combination of two methods. The outcome of classification of exoplanets based on the CDHS (Method 1) is tallied with another classification method which discriminates samples (exoplanets) into classes based on the features/attributes of the samples (Method 2). The similar outcome from both approaches (the exoplanets are classified into the same habitability class), markedly different in structure and methodology, fortifies the growing advocacy of using machine learning in astronomy.

The habitability score model considers four parameters/features, namely mass, radius, density and surface temperature of a planet extracted from the PHL-EC

[1] Department of Computer Science and Engineering, PESIT-BSC, Bangalore, E-mail: `snehanshusaha at pes.edu`

[2] Department of Information Science and Engineering, PESIT-BSC, Bangalore

[3] M. P. Birla Institute of Fundamental Research, Bangalore

[4] Indian Institute of Astrophysics, Bangalore

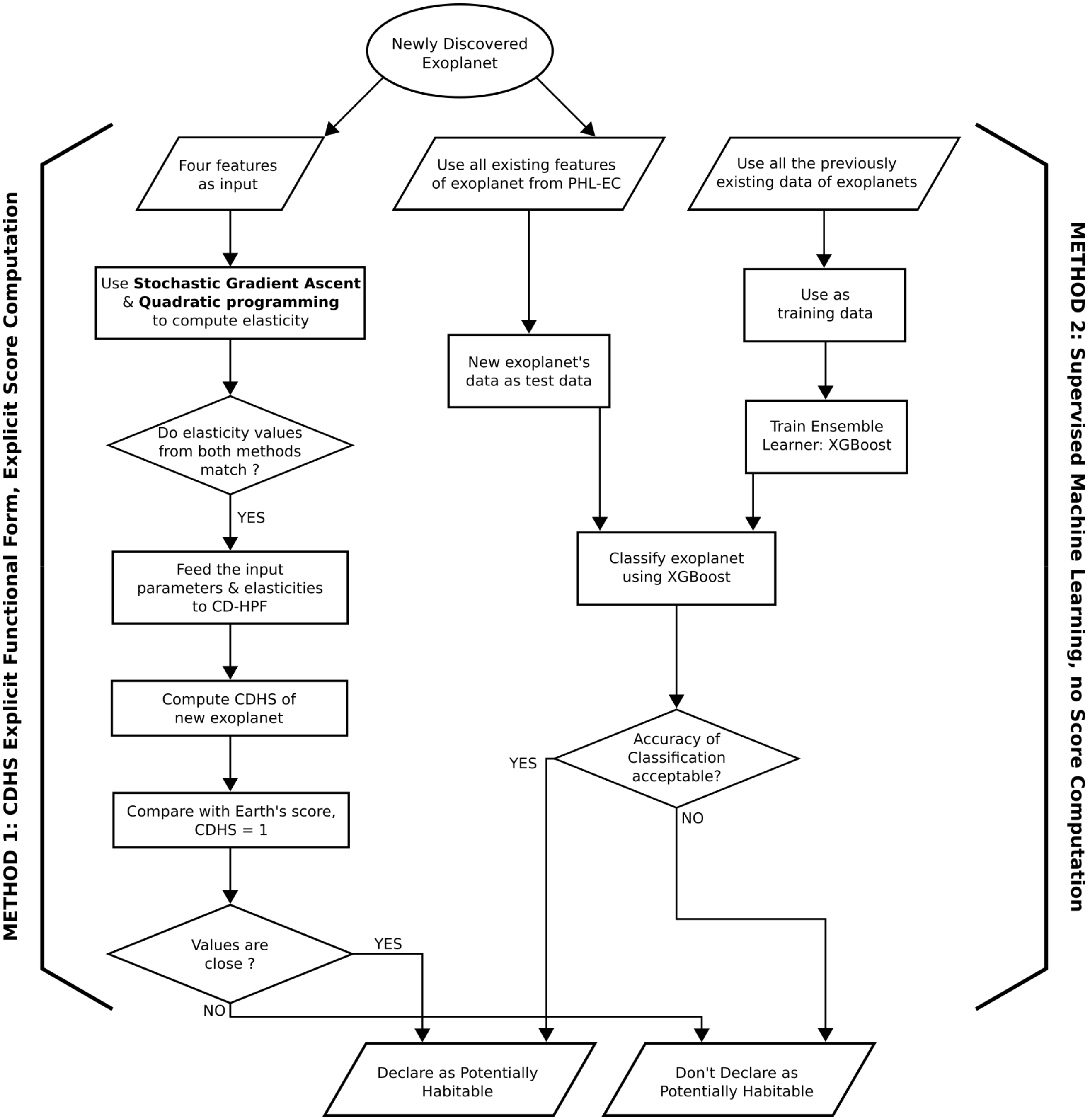
Newly Discovered Exoplanet
METHOD 1: CDHS Explicit Functional Form, Explicit Score Computation
METHOD 2: Supervised Machine Learning, no Score Computation
Four features as input
Use all existing features of exoplanet from PHL-EC
Use all the previously existing data of exoplanets
Use Stochastic Gradient Ascent & Quadratic programming to compute elasticity
Use as training data
New exoplanet's data as test data
Train Ensemble Learner: XGBoost
Do elasticity values from both methods match ?
YES
Feed the input parameters & elasticities to CD-HPF
Classify exoplanet using XGBoost
Compute CDHS of new exoplanet
Accuracy of Classification acceptable?
YES
NO
Compare with Earth's score, CDHS = 1
Values are close ?
YES
NO
Declare as Potentially Habitable
Don't Declare as Potentially Habitable

Fig. 1: The convergence of two different approaches in the investigation of potential habitability. The outcome of the explicit scoring scheme placed Proxima b in the “Earth League”, which is synonymous to being classified as potentially habitable.

(Exoplanet Catalog hosted by the Planetary Habitability Laboratory (PHL))[1]. Though the catalog contains 68 observed and derived stellar and planetary parameters, we have currently considered only four for the CDHS model. CDHS does encounter problems commonly found in convex functional modeling, such as scalability and curvature violation. We show here that the CDHS model is scalable, i.e. capable of accommodating more parameters (see Section IV on model scalability, and Section 3 in supplementary file (Proof of model scalability), [12] for the proof of the theorem). Therefore, we may use more parameters in future to compute the CDHS. The problem of curvature violation is tackled in Sec.IV later in the paper.

PHL classifies all discovered exoplanets into five categories based on their thermal characteristics: non-habitable, and potentially habitable: psychroplanet, mesoplanet, thermoplanet, and hypopsychroplanet. Proxima b and the TRAPPIST-1 system are amongst the recent additions to the catalog with recorded features. Here, we employ a non-metric classifier to predict the class label of the recently discovered exoplanets. We compute the accuracy of our classification method, and aim to reconcile the result with the habitability score of the recently discovered exoplanets, which may suggest its proximity to the "Earth League". We call this an investigation in the optimistic determination of potential habitability. The hypothesis is the following: a machine learning-based classification method, known as boosted trees, classifies exoplanets and returns some with the class by mining the features present in the PHL-EC (Method 2 in Figure 1). This process is independent of computing the explicit habitability score for recently developed exoplanets (*aka* Method 1 in Figure 1), and indicates habitability class by learning attributes from the catalog. This implicit method should match the outcome suggested by the CDHS. In other words, the class label of exoplanets predicted by the implicit method (Method 2) should correspond to the appropriate CDHS of those exoplanets. This is demonstrated in Table VII.

The second approach is based on XGBoost – a statistical machine-learning classification method used for supervised learning problems, where the training data with multiple features are used to predict a target variable. Authors intend to test whether the two different approaches to investigate the habitability of Proxima b and the TRAPPIST-1 planets, analytical and statistical, converge with a reasonable degree of confidence. The paper is split into two layers. The first layer considers Proxima b as a test case, and the latter layer applies our methods on the Trappist-1 system. Trappist-1 was discovered later than Proxima b and, hence the section Sec. VI was added.

The paper is organized as follows. Sections II, III, IV, and V elaborate the theory and methods and discuss the implications on Proxima b as a test case (Layer 1). Section VI forms layer 2, focusing on the Trappist-1 system. Section VII elaborates the overall efficacy of our approach applied to Proxima b and Trappist-1 system. Appendices present the theory in detail.

II. Analytical Approach via CDHS

We begin by discussing the key elements of the analytical approach. The parameters of Proxima b were extracted from the PHL-EC: minimal mass 1.27 EU, radius 1.12 EU, density 0.9 EU, surface temperature 262.1 K, and escape velocity 1.06 EU, where EU is the Earth Units. Its Earth Similarity Index (ESI), estimated using a simplified version[2] of the ESI, is 0.87. By definition, ESI range is from 0 (totally dissimilar to Earth) to 1 (identical to Earth), only planets with ESI $\geq$ 0.8 are considered to be Earth-like.

A. Earth Similarity Index

In general, the ESI value of any exoplanet's planetary property is calculated using the following expression [6],

$$ESI_x = \left(1 - \left|\frac{x - x_0}{x + x_0}\right|\right)^{w_x} , \tag{1}$$

where x is a planetary property – radius, surface temperature, density, or escape velocity, x_0 is the Earth's reference value for that parameter, i.e. 1 EU, 288 K, 1 EU and 1 EU, respectively, and w_x is the weighted exponent for that parameter. After calculating ESI for each parameter by Eq. 1, the global ESI is found by taking the geometric mean (G.M.) of all four ESI$_x$,

$$ESI = \left(\prod_{x=1}^{n} ESI_x\right)^{\frac{1}{n}} . \tag{2}$$

The problem in using Eq. 2 to obtain the global ESI is that sometimes there no available data to obtain all input parameters, such as in the case of Proxima b – only its mass and the distance from the star are known. Due to that, a simplified expression was proposed by the PHL for ESI calculation in terms of only radius and stellar flux,

$$ESI = 1 - \sqrt{\frac{1}{2}\left(\frac{R - R_0}{R + R_0}\right)^2 + \left(\frac{S - S_0}{S + S_0}\right)^2} , \tag{3}$$

where R and S represent radius and stellar flux of a planet, and R_0 and S_0 are the reference values for the Earth. Using 1.12 EU for the radius and 0.700522 EU for the stellar flux, we obtain ESI = 0.8692. It is worth mentioning that once we know one observable – the mass – other planetary parameters used in the ESI computation (radius, density and escape velocity) can be calculated based on certain assumptions. For example, the small mass of Proxima b suggests a rocky

[1] The latest updated (May 2017) dataset can be downloaded from the PHL website: http://phl.upr.edu/projects/habitable-exoplanets-catalog/data/database

[2] http://phl.upr.edu/projects/earth-similarity-index-esi

composition. However, since 1.27 EU is only a low limit on mass, it is still possible that its radius exceeds 1.5 – 1.6 EU, which would make Proxima b, not rocky [7]. In the PHL-EC, its radius is estimated using the mass-radius relationship

$$R = \begin{cases} M^{0.3} & M \leq 1 \\ M^{0.5} & 1 \leq M < 200 \\ (22.6)\, M^{(-0.0886)} & M \geq 200 \end{cases} \tag{4}$$

Since Proxima b mass is 1.27 EU, the radius is $R = M^{0.5} \equiv 1.12$ EU. Accordingly, the escape velocity was calculated by $V_e = \sqrt{2GM/R} \equiv 1.065$ (EU), and the density by the usual $D = 3M/4\pi R^3 \equiv 0.904$ (EU) formula. If we use all four parameters provided in the catalog, the global ESI becomes 0.9088.

B. Cobb Douglas Habitability Score (CDHS)

We have proposed the new model of the habitability score in [1] using a convex optimization approach [8]. In this model, the Cobb Douglas function [9] is reformulated as Cobb-Douglas habitability production function (CD-HPF) to compute the habitability score of an exoplanet,

$$\mathbb{Y} = f\left(R, D, T_s, V_e\right) = K\left(R\right)^{\alpha} \cdot \left(D\right)^{\beta} \cdot \left(T_s\right)^{\gamma} \cdot \left(V_e\right)^{\delta} \tag{5}$$

where the same planetary parameters are used – radius R, density D, surface temperature T_s, and escape velocity V_e. $\mathbb{Y}$ is the habitability score CDHS, and f is defined as CD-HPF [3]. The goal is to maximize the score, $\mathbb{Y}$, where the elasticity values of each parameter are subject to the condition $\alpha + \beta + \gamma + \delta < 1$. Note that the interior CDHS$_i$, denoted by $Y1$, is calculated using radius R and density D, while the surface CDHS$_s$, denoted by $Y2$, is calculated using surface temperature T_s and escape velocity V_e. The objective is to find elasticity value that produces the optimal habitability score for the exoplanet, i.e. to find $Y_1 = \max_{\alpha,\beta} Y(R, D)$ such that, $\alpha > 0$, $\beta > 0$ and $\alpha + \beta \leq 1$. Similarly, we need to find $Y_2 = \max_{\gamma,\delta} Y(T, V_e)$ such that $\gamma > 0$, $\delta > 0$ and $\delta + \gamma \leq 1$. Elasticity values are obtained by a computationally fast Stochastic Gradient Ascent (SGA) algorithm described in Sec. III-A. We calculate CDHS score for the constraints known as returns to scale: Constant Return to Scale (CRS) and Decreasing Return to Scale (DRS) (for details, see [1]). Note that $\alpha + \beta < 1$ is the DRS condition for elasticity, which may be scaled to $\alpha_1 + \alpha_2 + \ldots + \alpha_n < 1$. Analogously, $\delta + \gamma < 1$ is the DRS condition for elasticity which may be scaled to $\delta_1 + \delta_2 + \ldots + \delta_n < 1$.

As Proxima b is considered an Earth-like planet, we endeavored to cross-match the observation via the method explained in the previous section. The analysis of CDHS will help to explore how this method can be effectively used for newly discovered planets. The eventual classification of any exoplanet is accomplished by using the proximity of CDHS of that planet to the Earth, with additional constraints imposed on the algorithm termed "probabilistic herding". The algorithm works by taking a set of values in the neighborhood of 1 (CDHS of Earth). A threshold of 1 implies that CDHS value between 1 and 2 is acceptable for membership in the "Earth-League", pending fulfillment of further conditions. For example, the CDHS of the most potentially habitable planet before Proxima b, Kepler-186 f, is 1.086 (the closest to the Earth's value), though its ESI is only 0.64. While another PHP – GJ-163 c has the farthest score (1.754) from 1; and though its ESI is 0.72, it may not be even a rocky planet as its radius can be between 1.8 to 2.4 EU, which is not good for a rocky composition theory (see e.g. [7]).

Sometimes, values of certain parameters are not available in the catalog (e.g. for 11 planets PHL-EC does not provide surface temperatures). In machine learning, the missing values can be imputed by using association rules, in particular, the *rule-based learning.* We have devised an algorithm based on [10] and [11] to impute missing values, the details of which are explained in supplementary file [12].

C. CDHS calculation using radius, density, escape velocity and surface temperature

Using the values of the parameters from the PHL-EC, we calculated CDHS score for the CRS and DRS cases, and obtained optimal elasticity and maximum CDHS value. The CDHS values in CRS and DRS cases were 1.083 and 1.095, respectively. The degree/extent of closeness is explained in [1] in great detail.

D. CDHS calculation using stellar flux and radius

Following the simplified version of the ESI (Eq. 3), we repeated the CDHS computation using only radius and stellar flux (1.12 EU and 0.700522 EU, respectively). From the scaled down version of Eq. 5, we obtain CDHS$_{CRS}$ and CDHS$_{DRS}$ as 1.083 and 1.095, respectively. These values confirm the robustness of the method used to compute CDHS and validate the claim that Proxima b falls into the "Earth-League" category.

E. CDHS calculation using stellar flux and mass

The habitability score requires the use of available physical parameters, such as radius, or mass, and temperature, and the number of parameters is not extremely restrictive. As long as we have the measure of the interior similarity – the extent to which a planet has a rocky interior, and exterior similarity – the location in the HZ or the favorable range of surface temperatures, we can reduce (or increase) the number of parameters. Since radius is calculated from an observable parameter – mass, we decided to use the mass directly in the calculation, obtaining CDHS$_{DRS}$ as 1.168 and CDHS$_{CRS}$ as 1.196. The CDHS achieved using radius and stellar flux (previous subsection) and the CDHS achieved using mass and stellar flux have the same values.

Remark: *Does this imply that stellar flux and planet mass are enough to compute the habitability score as*

[3] Elasticities K, α, β, γ and δ need to be estimated

defined by our model? It cannot be confirmed until enough number of clean data samples are obtained containing the four parameters used in the original ESI and CDHS formulation. We plan to perform a full-scale dimensionality analysis as future work

The values of ESI and CDHS using different methods are summarized in Table I.

TABLE I: ESI and CDHS values calculated for different parameters

Parameters Used	ESI	CDHS$_{CRS}$	CDHS$_{DRS}$
R, D, T_s, V_e	0.9088	1.083	1.095
Stellar Flux, R	0.869	1.196	1.168
Stellar Flux, M	0.849	1.196	1.167

NOTE: The nicety in the result, i.e. little difference in the values of CDHS, is due to the flexibility of the functional form in the model proposed in [13], and the computation of the elasticities by the Stochastic Gradient Ascent method described in the next section. Using this method led to the fast convergence of the elasticities. Proxima b passed the scrutiny and is classified as a member of the "Earth League".

III. Elasticity computation: Stochastic Gradient Ascent (SGA)

[1] used a library function **fmincon** to compute the elasticity values. Here, we have implemented a more efficient algorithm to perform the same task. This was done for two reasons: to be able to break free from the inbuilt library functions, and to devise a sensitive method which would mitigate oscillatory nature of Newton-like methods around the local minima/maxima. There are many methods which use gradient search, including the one proposed by Isaak Newton. Although theoretically sound, algorithmic implementations of most of these methods face convergence issues in real time due to the oscillatory nature. We have employed a modified version of the descent, an SGA algorithm, to calculate the optimum CDHS and the elasticities for mass, radius, density and escape velocity (Eq. 5 in Sec. II-B). As opposed to the conventional Gradient Ascent/Descent method, where the gradient is computed only once, stochastic version recomputes the gradient for each iteration and updates the elasticity values. Theoretical convergence, guaranteed otherwise in the conventional method, is sometimes slow to achieve though. Stochastic variant of the method speeds up the convergence, justifying its use in the context of the problem (the size of data, i.e. the number of discovered exoplanets, is increasing every day).

Output elasticity (α, β, γ or δ) of Cobb-Douglas habitability function is the accentual change in the output in response to a change in the levels any of the inputs. Accuracy in elasticity values is crucial in deciding the right combination for the optimal CDHS, where different approaches are analyzed before arriving at final decision. In the next subsections, we show how the elasticities were computed on the example of α and β. Once they are computed, we repeat the procedure to compute other elasticities, γ and δ.

A. Computing Elasticities via Gradient Ascent

Gradient Ascent is an optimization algorithm used for finding the local maximum of a function. Given a scalar function $F(x)$, gradient ascent finds the $\max_x F(x)$ by following the slope of the function. This algorithm selects initial values for the parameter x and iterates to find the new values of x which maximizes $F(x)$ (here CDHS). Maximum of a function $F(x)$ is computed by iterating through the following step,

$$x_{n+1} \leftarrow x_n + \chi \frac{\partial F}{\partial x}\,, \tag{6}$$

where x_n is an initial value of x, x_{n+1} the new value of x, $\frac{\partial F}{\partial x}$ is the slope of function $Y = F(x)$ and χ denotes the step size, which is greater than 0 and forces the algorithm to make a small jump (descent or ascent algorithms are trained to make small jumps in the direction of the new update). Stochastic variant thus mitigates the oscillating nature of the global optima – a frequent malaise in the conventional Gradient Ascent/Descent and Newton-like methods, such as **fmincon** used in [1]. At this point of time, without further evidence of recorded/measured parameters, it may not be prudent to scale up the CD-HPF model by including more parameters other than the ones used by either ESI or our model. But if it ever becomes a necessity (to utilize more than the four parameters), the algorithm will come in handy and multiple optimal elasticity values may be computed fairly easily.

B. Computing Elasticities via Constrained Optimization

Let the assumed parametric form be $\log(y) = \log(K) + \alpha\log(S) + \beta\log(P)$[4]. Consider a set of data points,

$$\begin{array}{ccccccc} \ln(y_1) & = & K' & + & \alpha S'_1 & + & \beta P'_1 \\ \vdots & & \vdots & & \vdots & & \vdots \\ \ln(y_N) & = & K' & + & \alpha S'_N & + & \beta P'_N \end{array} \tag{7}$$

where $K' = \log(K)$, $S'_i = \log(S'_i)$ and $P'_i = \log(P'_i)$. If $N > 3$, this is an over-determined system, where one possibility to solve it is to apply a least squares method. Additionally, if there are constraints on the variables (the parameters to be solved for), this can be posed as a constrained optimization problem. These two cases are discussed below.

No constraints: This is an ordinary least squares solution. The system is in the form $y = Ax$, where

$$x = \begin{bmatrix} K' & \alpha & \beta \end{bmatrix}^T, \quad y = \begin{bmatrix} y_1 \\ \cdot \\ \cdot \\ y_N \end{bmatrix}, \tag{8}$$

[4] This is a logarithmic transformation of the standard CDHS model (which has the exponential form).

and

$$A = \begin{bmatrix} 1 & S_1' & P_1' \\ & ... & \\ 1 & S_N' & P_N' \end{bmatrix} . \tag{9}$$

The least squares solution for x is the solution that minimizes

$$(y - Ax)^T (y - Ax) . \tag{10}$$

It is well known that the least squares solution to Eq. (8) is the solution to the system $A^T y = A^T Ax$, i.e. $x = (A^T A)^{-1} A^T y$. In *Matlab*, the least squares solution to the overdetermined system $y = Ax$ can be obtained by $x = A \backslash y$. Table II presents the results of least squares (**no constraints**) obtained for the elasticity values after performing the least square fitting, while Table III displays the results obtained for the elasticity values after performing the constrained least square fitting; in Table IV, the values of CRS and DRS from quadratic programming have been enunciated.

TABLE II: Elasticity values for IRS, CRS & DRS cases after performing the least square test (**no constraints**): elasticities α and β satisfy the theorem $\alpha+\beta < 1, \alpha+\beta = 1$, and $\alpha + \beta > 1$ for DRS, CRS and IRS, respectively, and match the values reported previously in [1].

	IRS	CRS	DRS
α	1.799998	0.900000	0.799998
β	0.100001	0.100000	0.099999

Constraints on parameters: This results in a constrained optimization problem. The objective function to be minimized (maximized) is still the same, namely,

$$(y - Ax)^T (y - Ax) . \tag{11}$$

This is a quadratic form in x. If the constraints are linear in x, then the resulting constrained optimization problem is a quadratic program (QP). A standard form of a QP is

$$\max x^T Hx + f^T x , \tag{12}$$

such that

$$\begin{aligned} Cx &\leq b ; && \text{Inequality constraint} \\ C_{\text{eq}} x &= b_{\text{eq}} ; && \text{Equality constraint.} \end{aligned}$$

Suppose the constraints are $\alpha, \beta > 0$ and $\alpha + \beta \leq 1$. The QP can be written as (neglecting the constant term $y^T y$)

$$\max x^T (A^T A) x - 2y^T Ax , \tag{13}$$

such that

$$\begin{cases} \alpha > 0 , \\ \beta > 0 , \\ \alpha + \beta \leq 1 . \end{cases} \tag{14}$$

For the standard form as given in Eq. (12), Eqs. (13) and (14) can be represented by rewriting the objective function as:

$$x^T Hx + f^T x , \tag{15}$$

where

$$H = A^T A \text{ and } f = -2A^T y . \tag{16}$$

The inequality constraints can be specified as

$$C = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 1 \end{bmatrix} , \text{ and } b = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} . \tag{17}$$

In *Matlab*, the QP can be solved using the function **quadprog**. The results in Table III were obtained by conducting quadratic programming.

TABLE III: Elasticity values for IRS, CRS & DRS cases after performing the least square test (**with constraints**): elasticity values α and β satisfy the theorem $\alpha + \beta < 1, \alpha + \beta = 1$, and $\alpha + \beta > 1$ for DRS, CRS and IRS, respectively, and match the values reported previously [1].

	IRS	CRS	DRS
α	1.799998	0.900000	0.799998
β	0.100001	0.100000	0.099999

Using active set: We have conducted the experiment using active learning technique. This framework is best suited in our case as it can be applied to different performance targets and all types of classifications. The traditional active-set method is divided into two steps, focusing on feasibility and optimality, in that order. Instead of an “ad-hoc” start, active set methods bank on a good “initiator” estimate of the optimal active set. This is well suited for a sequence of quadratic programs to be solved, which is what our constrained optimization problem needs. Active set gave best results out of all the three algorithms which suffices our argument.

TABLE IV: Results of quadratic programming by using the active-set learning. Exact match with SGA results and Method 1, which satisfy the conditions of CRS, DRS & IRS i.e. elasticity values α and β satisfy the theorem $\alpha + \beta = 1; \alpha + \beta < 1; \alpha + \beta > 1$ and match the values reported in [1].

	CRS	DRS
K	1	1
α	0.9000	0.8000
β	0.1000	0.1000

By solving the described constrained QP, we find that the result satisfy the condition $\alpha + \beta \leq 1$ for both CRS and DRS cases, and the condition $\alpha + \beta \geq 1$ for the IRS case. Elasticities α, β and K from both computations are very close, supporting our choice of α, β and K. Identical results are observed for elasticities γ and δ for the surface CDHS. Obtained elasticity values used for computing CDHS are $\alpha = 0.8$ and $\beta = 0.1$ for DRS, and $\alpha = 0.9$ and $\beta = 0.1$ for CRS cases, respectively. The algorithms are repeated to compute γ and δ, and a convex combination of interior and surface CDHS is used to calculate the CDHS of Proxima b. The entire process, including classification of all exoplanets post-habitability score computation, is summarized in Appendices.

IV. Model Scalability

In paper [1], we have shown the theoretical guarantee regarding the conditions on elasticity. However, the scalability of the model (scalability of the theoretical guarantee) depends on the fact that the conditions of global maxima continue to hold even if the number of input parameters increase. In addition, the theoretical guarantee in some cases [8] tends to relax when an arbitrary parameter is added to the model. This happens due to curvature violation of the functional form. In other words, if eccentricity (say) is added as one of the input parameters along with surface temperature, density, radius and mass, there needs to be a mathematical guarantee that the conditions on elasticity should scale in the same fashion. This has been illustrated previously via computer simulation. However, there needs to be a theoretical result fortifying the intuitive understanding of the proposed model and the scoring scheme – the CDHS. We define the theorem which lays the foundation for model scalability in the event any parameter is added to the existing model, already accommodating an arbitrary number of parameters. If the conditions of elasticity for a global maxima hold for a fixed set of input parameters (say, n), it will continue to hold when the number of parameters is increased by 1 (say, $n+1$). This is an inductive approach, non-traditional but powerful! The proof (given in Section 3 of [12]) is based on the principle of mathematical induction.

Theorem: If global maxima for CDHS, i.e.

$$\log(Y) = \frac{1}{1-\sum\limits_{i=1}^{n}\alpha_i}\log\left\{k\prod_{i=1}^{n}\left(\frac{x_i p}{w_i}\right)^{\alpha_i}\right\} \tag{18}$$

holds, then the same condition for the global maxima will continue to hold if an additional input parameter is inserted in the habitability function CD-HPF, i.e., if

$$\log(Y_{\text{new}}) = \frac{1}{1-\sum\limits_{i=1}^{n+1}\alpha_i}\log\left\{k\prod_{i=1}^{n+1}\left(\frac{x_i p}{w_i}\right)^{\alpha_i}\right\}$$

holds as well. Further, it follows that the elasticity condition for DRS for $n+1$ parameters is true, i.e. $1-\sum\limits_{i=1}^{m+1}\alpha_i > 0$, if the elasticity condition for DRS for n parameters, i.e. $1-\sum\limits_{i=1}^{m}\alpha_i > 0$, holds.[5]

We have investigated the habitability of newly discovered exoplanets via CDHS (Method 1). CDHS leads to a classification scheme (Algorithm 3 in Section 4 of the supplementary material) and depends on computing the habitability score of discovered exoplanets. However, the classification problem doesn't have to rely on having numerical values of the response variable of samples under classification. Instead, the hidden relationship between samples may be discovered by construction of the decision rules connecting the feature values of the samples. In the next section, we will explore the habitability classification problem from a supervised learning perspective (Method 2), where a collection of labeled data (exoplanets from the PHL-EC) is used as training set, and Proxima b is used as the test data. We train the machine to learn the features associated with the training and test data and identify the class label of the test data via machine classification algorithm, known as XGBoost. The goal, as stated earlier, is to test the ability of the algorithm to label Proxima b in the "Earth League" with a reasonably high accuracy, thereby establishing the strong correlation between the two different approaches.

V. Classification of Proxima Centauri b via non-functional form: XGBoost, a feature-based learning and classification method

Here we illustrate a method by which the high habitability score of Proxima b may be predicted by using class labels and features from the PHL-EC. The method XGBoost (eXtreme Gradient Boosting) is a non-metric classifier, and a fairly recent addition to the suite of machine learning algorithms [14]. Non-metric classifiers are applied in scenarios where there are no definitive notions of similarity between feature vectors.

A typical machine-learning problem processes input data and combines that with the learning algorithm to produce a model as output. Learning implies recognizing complex patterns and making intelligent decisions based on data. The machine comes up with its own prediction rule, based on which a previously unobserved sample would be classified as a certain type, meso or psychroplanets for example, with a reasonable accuracy.

In order to appropriately apply a method (including preprocessing and classification), a thorough study of the nature of the data should be done; this includes understanding the number of samples in each class, the separability of the data, etc. Depending on the nature of the data, appropriate preprocessing and post processing (if needed) methods should be determined along with the right kind of classifier for the task.

A. Understanding the data to be classified

The PHL-EC dataset contains more than 3500 samples and is growing steadily: from 1904 samples in November 2015 to 3635 samples at the time of writing [6]. We have considered 51 features of the data for classifying, and

[5] Remark:The habitability score CDHS is computed using four parameters: R, D,T_s and V_e. If a new parameter from the PHL-EC needs to be added to the CD-HPF, it is important to know if the conditions of global maxima for habitability still holds. The above theorem validates our superposition conclusively.

[6] These numbers vary over time. We need these samples to train the classifier, so that it can classify new additions, such as e.g. Proxima b or Trappist-1 planets.

have eliminated the ones that are unimportant for classification, such as the name of the parent star (*S.Name*), the name of the planet (*P.Name*), etc. In the dataset PHL-EC, planets are already segregated into five classes based on their surface thermal properties:

1) **Non-Habitable**: planets that do not have thermal properties required to sustain life.
2) **Mesoplanet**: planets with a mean global surface temperature between 0°C and 50°C – a necessary condition for complex terrestrial life. These are generally referred to as Earth-like planets.
3) **Psychroplanet**: planets with mean global surface temperature between −15°C and +10°C – somewhat colder than the optimal temperature for the sustenance of terrestrial life.
4) **Thermoplanet**: planets with the temperature in the range of 50°C – 100°C – warmer than the temperature range suited for most terrestrial life.
5) **Hypopsychroplanets**: planets with temperature below −50°C. These planets are too cold for the survival of most terrestrial life.

Out of these, the classes of hypopsychroplanet and thermoplanet have too few samples (only two planets each) and hence are not useful for the analysis. The classification was performed on remaining three classes: psychroplanet, mesoplanet and non-habitable.

A planet having characteristics suitable for inhabitation is still a rare occurrence; naturally, most of the samples in the dataset belong to the class of non-habitable planets (3592 out of 3635). From a data analytic point of view, this is a *data bias* and can lead to overfitting, i.e., when a classifier becomes overly complex and extremely sensitive to the nuances in the data. Overfitting is a problem that needs to be dealt with carefully and not be overlooked as an administrative task. In a dataset such as the PHL-EC, where the number of samples belonging to one class is over a thousand times the total number of samples belonging to all the other classes, just reporting the numeric accuracy obtained by directly feeding the data to train a classifier would be an incorrect methodology.

To counter the potential problems due to the dominance by a single class, we used *artificially balanced* datasets by considering random samples from the classes of non-habitable and mesoplanets with the total number of samples belonging to one class being equal to the number of samples in the psychroplanet class, as it has the least number of samples. Then this balanced dataset was divided in a ratio 9:4 (we found this to be the best ratio), where the larger portion was that of the training set. This cycle of balancing the data set artificially, dividing it, training and testing a classifier was performed multiple times, and the mean accuracy of all the trials was considered to be representative of the potential of a classifier. By artificial balancing, the reported accuracies are also more reliable than without balancing.

We have applied a powerful ensemble classification for the task described above. Boosting refers to the method of combining the results from a set of *weak learners* to produce a *strong* prediction. Generally, a weak learner's performance is close to a random guess. A weak learner divides the job of a single predictor across many weak predictor functions, and optimally combines the votes from all smaller predictors. This helps enhancing the overall prediction accuracy.

XGBoost is a tool developed by utilizing these boosting principles [14]. XGBoost combines a large number of regression trees with a small learning rate. As regression can be used to model classifiers: here the word *regression* may refer to logistic or soft-max regression for the task of classification. XGBoost uses an ensemble of decision trees. We describe the detailed working principle in Section 6 (XGBoost: An Exploration of Machine Learning based Classification) contained in [12].

B. Classification of Data

As a first step, data from PHL-EC is pre-processed (the authors have tried to tackle the missing values by taking mean for continuous-valued attribute, and mode for categorical attributes). Certain attributes from the database, namely *P.NameKepler, S.nameHD, S.nameHid, S.constellation, S.type, P.SPH, P.interiorESI, P.surfaceESI, P.disc.method, P.disc.year, P.maxmass, P.minmass, P.inclination* and *Habmoon* were removed as these attributes do not contribute to the nature of classification of habitability of a planet. Though individual ESI values (and planetary mass) do contribute to habitability determination, because the data set directly provides total value of *P.ESI* – the global ESI of the planet, these features were neglected. Following this, classification algorithms were applied on the processed data set, where in total 49 features were used.

The pursuit of finding the appropriate classification method for any classification problem requires a lot of experimentation and analysis of the nature of the data. We performed a convex hull test to understand the nature of the data, and found that the data is not linearly separable. Hence, classifiers like SVM (Support Vector Machines), k-NN (k Nearest Neighbors) and LDA (Linear Discriminant Analysis) are not expected to perform well. All these classifiers were tried as candidates for classification; as expected from the convex hull test, they did not perform well. This motivated our choice of Tree-based classifiers and, more specifically, of the XGBoost.

After understanding the nature and separability of the data, a more suitable approach was developed. In this whole classification process, the PHL-EC data set had 3411 entries, from the data set obtained in November 2016: 24 entries belonging to class mesoplanet, 13 entries belonging to class psychroplanet, and 3374 entries belonging to class non-habitable [7]. The number of items in this data set was significantly more than the older data set used to have. Hence, the artificial balancing method was modified. In the new balancing method, all 13 entries from psychroplanet class were considered in a smaller data set, and 13 random and unique entries from each of the other two classes were also considered. Thus, in this case, the number of entries in a smaller, artificially balanced data set was 39. Following this, each smaller data set was balanced in the ratio of 9:4 (training:testing) and 500 iterations of training and testing were performed on each such data set. 500 such data sets were framed for analysis. To sum it up, 2,50,000 iterations of training-testing were performed.

C. *Accuracy of algorithms used to Classify Proxima b*

XGBoost was used to classify the conservative and optimistic samples from the PHL-EC. The ROC (Receiver Operating Characteristic) curve obtained for this classification is shown in Figure 2. Each point on the ROC plot represents a sensitivity/specificity pair which corresponds to a particular decision threshold. Sensitivity, or recall, is the the proportion of positive tuples that are accurately identified, and specificity is the the proportion of negative tuples that are correctly identified. A test with non-overlapping classes has an ROC plot that passes through the upper left corner (100% sensitivity, 100% specificity). Therefore, the closer the ROC plot is to the upper left corner, the higher the overall accuracy of the test [15].

The accuracy of the XGBoost algorithm used to classify Proxima b was 100%, i.e. there were no false positives or false negatives in its classification. The method of classification was to select Proxima b as the training set and the remaining samples in the catalog as the test set (subject to artificial balancing by under-sampling the non-habitable class).

However, in cases where a dataset exhibits a data bias towards one class, the F-score test statistic is more representative than the accuracy of a classifier. It is used to analyze whether a classifier is able to achieve both high precision and high recall simultaneously (for details on precision, recall, F-score, ROC curves, etc., see [16], [17]). The values for precision and recall were calculated to gauge the goodness of the classifier. This was done but considering the psychroplanet and mesoplanet classes as the positive classes one at a time. With respect to the psychroplanet class, the calculated F-score was 0.94, precision was 0.95, and recall was 0.93. With respect to the class of mesoplanets, the F-score was 0.95, precision was 0.93, and recall was 0.97. Using the XGBoost classifier, the class-belongingness of Proxima b to the class of psychroplanets was estimated to be 100%. This is indeed the true class of Proxima b. We can thus say that XGBoost performs the classification remarkably well!

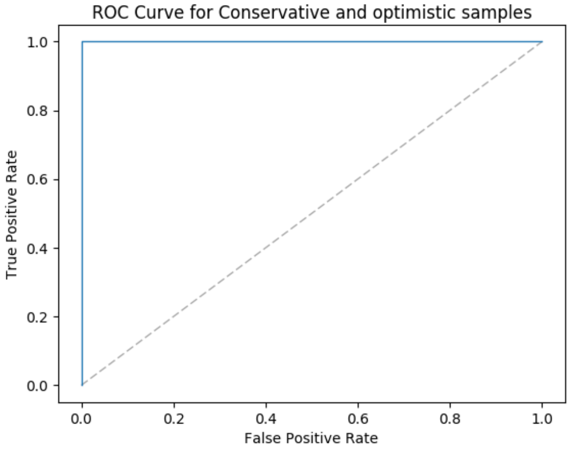

Fig. 2: ROC curve (blue line) for the classification of the conservative and optimistic samples: these samples include Proxima b and Kepler 186-f. The dotted line is the ROC for the case when the performance of the binary classifier is the same as a perfectly random guess. A good numeric representation of ROC curves is the percentage area of the coordinates which falls under the curve (AUC). Here, the AUC is 100%, which indicates a perfect performance of the classifier.

VI. The TRAPPIST-1 system

This section focuses on the application of the same algorithms and methods discussed in previous sections for the classification and CDHS computation of the planets in the TRAPPIST-1 system.

The discovery of the TRAPPIST-1 system has caught the attention of the entire astronomy community recently [5], [18]. TRAPPIST-1 is an ultra-cool dwarf, detected by the 2MASS Sky Survey. Following this study, a series of papers were published [19] by various researchers working on the exoplanets, and for a good reason: all seven planets in the TRAPPIST-1 system are likely Earth-sized and rocky, with the estimated low equilibrium temperatures — due to the exceptionally low stellar luminosity (1/1000th of the Sun), the insolation on the planets is equivalent to the insolation on the terrestrial group, thus allowing the possibility of liquid water on the surface. Three of the planets are within in the stellar habitable zone. Though all planets are most probably tidally locked with the parent star, water could still exist even on the innermost planets [5].

We have used Cobb-Douglas habitability model to compute CDHS of TRAPPIST-1 planets, as well as classify them using K-NN classifier. The result of the experiment is shown in Table V. We conclude from the result (a consequence of Method 1, Figure 1) that the planets are in the “Earth-League”.

[7] These are different from the numbers reported above. This is natural as discovery of exoplanets is a continuous process. Please note, as and when the catalog is updated, we update the training population as well.

TABLE V: Observed and calculated parameters of TRAPPIST-1 planets. Physical parameters are given in Earth units (EU). Computed habitability score and class of the planets belonging to planets system satisfy the threshold condition (Algorithm 3). The outcome of the machine learning algorithm fortifies the experimental findings. Please refer to Figure 1 for visual illustration. Class label 6 implies most likely habitable exoplanets in a probabilistic sense (see Earth-League candidate selection algorithm in Section 4 (Algorithm for Habitability candidacy classification) of [12] and [1].)

Planet	Mass	Radius	Mean Insolation	Mean $T_s(K)$	HZ	CDHS$_{DRS}$	CDHS$_{CRS}$	Class	$P.ESI$
b	0.86	1.09	4.2	396.5	no	1.0318	1.0410	5	0.56
c	1.38	1.06	2.25	347.9	no	1.14084	1.1589	5	0.73
d	0.41	0.77	1.13	292.4	no	0.9642	0.8870	5	0.9
e	0.64	0.92	0.65	260.4	yes	0.9722	0.9093	6	0.85
f	0.67	1.04	0.38	229.7	yes	0.9803	0.9826	6	0.68
g	1.34	1.13	0.26	216.1	yes	1.0951	1.1085	6	0.58
h	0.35	0.75	0.14	181.8	no	0.9511	0.8025	5	0.45

VII. Discussion and Conclusion

The discovery of Proxima b was announced on 24th August 2016 [2]. Proxima b is, at least, 1.3 times heavier than Earth. According to the PHL-EC, its radius is 1.12 EU, density is 0.9 EU, surface temperature is 262.1 K, and escape velocity is 1.06 EU. These attributes are close to those of the Earth, hence, there are plausible reasons to believe that Proxima b may be a habitable planet. In the PHL-EC data set, Proxima b is classified as a psychroplanet. We have computed the habitability score CDHS of Proxima b using different combination of planetary parameters: radius, density, escape velocity and surface temperature; only surface temperature and radius; stellar flux and radius; and stellar flux and mass. According to our classification algorithm, Proxima b falls in the Earth's class – "Earth-League" [1]. Its habitability "floor function value" is 1, and the difference between its CDHS and the Earth's CDHS is within the acceptable threshold of 1, as discussed in the published paper. The classification model XGBoost was used in this work to classify Proxima b with an accuracy of 100% (no false negatives or false positives). The accuracy results provide evidence of the strength of the model to automatically label and classify newly discovered exoplanets, such as Proxima b, or TRAPPIST-1 planets) in this case.

TABLE VI: Habitability score of Proxima b and Kepler-186 f, two planets most potentially habitable planets before the TRAPPIST-1 system discovery. Earth's ESI and CDHS are both 1 and considered as the baseline. CDHS values of the exoplanets in this table and Table V are significantly closer to the baseline score (Earth's score) compared to the ESI.

Planet Name	CDHS (DRS)	CDHS (CRS)	$P.ESI$
Kepler 186 f	1.075074	1.086295	0.61
Proxima Centauri b	1.08297	1.095255	0.87

Our algorithm is emphatically exhibiting the validation of the potential habitability of Proxima b, matching with the PHL findings. The robustness of the formula and the solid theory behind the formulation are validated by the proximity of the scores computed for different cases.

We have worked on two ways of affirming the habitability score of a planet. Essentially, we answer two questions: "*Is this planet potentially habitable?*" and "*How potentially habitable is this planet?*". These two questions are like the two sides of the same coin. By performing classification, we can affirm if a planet is expected to be potentially habitable or not, and by computing the CDHS, we are basically assigning a number to every planet which reflects its potential habitability. By doing this, in the future, we can gain deeper insights to planet's characteristics, understand what range of scores of the CD-HPF implies which classes of habitability, approximate unobserved attributes of a planet, etc. As the volume of data in the PHL-EC catalog increases with time, a robust automated method must be in place to analyze the data quickly and in an efficient way. An automated system primarily serves two purposes. The first is that it reduces human error in computation. The second, it eradicates the subjectivity that arises when different people try to classify or judge any data sample (here, a planet). One researcher's appraisal of a data sample might not be the same as that of another when evaluated based on general characteristics. However, when an algorithm is used for this, the results will be the same, regardless of which computer the system is deployed on. Hence, the implication of a system like this is the standardization of classification and of multiple ways of evaluating the potential habitability of an exoplanet.

This system may be extended in the future to analyze how the CDHS correlate with the classes of habitability. As there are multiple classes, it would be interesting to see if the CDHS falls into certain ranges for each class. The convergence of the score, however, gives rise to the following questions:

1) Are only stellar flux/surface temperature and radius/mass enough to construct a reliable habitability score via machine learning?
2) Should a full-scale dimensionality reduction technique be employed (completely data-driven approach) in the future, to analyze the context and validate such a claim?

TABLE VII: Summary of results of both methods: samples which are labeled as Class 6, an indicator of potential habitability are also predicted as habitable: the outcome of both approaches matches. For example, Trappist 1-e which is labeled as psychroplanet by Method 2 with 100% accuracy is also in Class 6 (most likely habitable class, [1]) according to the classification method 1.

Exoplanet	Method 1: Explicit Score Calculation			Method 2: Classification using XGBoost	
	$CDHS_{DRS}$	$CDHS_{CRS}$	Class Category	Classification Accuracy (%)	Predicted Class
Kepler 186 f	1.075074	1.086295	class	–	–
Proxima b	1.08297	1.095255	class	100.0	psychroplanet
TRAPPIST-1 b	1.0318	1.0410	5	89.6	non-habitable
TRAPPIST-1 c	1.14084	1.1589	5	88.4	non-habitable
TRAPPIST-1 d	0.9642	0.8870	5	100.0	mesoplanet
TRAPPIST-1 e	0.9722	0.9093	6	100.0	psychroplanet
TRAPPIST-1 f	0.9803	0.9826	6	99.7	psychroplanet
TRAPPIST-1 g	1.0951	1.1085	6	82.3	psychroplanet
TRAPPIST-1 h	0.9511	0.8025	5	95.1	non-habitable

Context is critical in solving a problem as complex as determining the habitability of discovered exoplanets. We mention with great regard the advances and contributions made by Dirk Schulze-Makuch, Abel Mendez, and other researches working in this field. In contrast to the ESI metric, our approach is entirely data-driven and inspired by machine learning. Our methods and algorithm cross-match the observation that Proxima b falls in the Earth category and is potentially habitable. This is an ample testimony of the efficacy of the proposed work.

CD-HPF is a novel metric of defining habitability score for exoplanets. It needs to be noted that the authors perceive habitability as a probabilistic measure, or a measure with varying degrees of certainty. Therefore, the construction of different classes of habitability classes 1 to 6 is contemplated, corresponding to measures as "most likely to be habitable" as Class 6, to "least likely to be habitable" as Class 1. As a further illustration, classes 6 and 5 seem to represent the identical patterns in habitability, but they do not! Class 6 – the "Earth-League" – is different from Class 5 in the sense that it satisfies the additional conditions of thresholding and probabilistic herding and, therefore, ranks higher on the habitability score. This is in stark contrast to the binary definition of exoplanets being "habitable or non-habitable", and a deterministic perception of the problem itself. The approach therefore required classification methods that are part of machine learning techniques and convex optimization — a sub-domain strongly coupled with machine learning. Cobb-Douglas function and CDHS are used to determine habitability and the maximum habitability score of all exoplanets with confirmed surface temperatures in the PHL-EC. Global maxima are calculated theoretically and algorithmically for each exoplanet, exploiting intrinsic concavity of CD-HPF and ensuring no curvature violation. Computed scores are fed to the attribute enhanced K-NN algorithm — a novel classification method, used to classify the planets into different classes to determine how similar an exoplanet is to Earth. The authors would like to emphasize that, by using classical K-NN algorithm and not exploiting the probability of habitability criteria, the results obtained were pretty good, having 12 confirmed potentially habitable exoplanets in the "Earth League". We have created a web page (`https://habitabilitypes.wordpress.com/`) for this project to host all relevant data and results: sets, figures, animation video and a graphical abstract. The web page contains the full customized catalog of all confirmed exoplanets with class annotations and computed habitability scores. The catalog is built with the intention of further use in designing statistical experiments for the analysis of the correlation between habitability and the abundance of elements (this work is briefly outlined in [20]). It is a very important observation that our algorithm and method give rise to a score metric, CDHS, which is structurally similar to the PHI (Planetary Habitability Index; [6]) as a corollary in the CRS case (when the elasticities are assumed to be equal to each other). Both are geometric means of the input parameters considered for the respective models.

CD-HPF uses four parameters (radius, density, escape velocity and surface temperature) to compute habitability score, which by themselves are not sufficient to determine habitability of exoplanets. Sometimes, there is a missing data in the catalogs, such as the case with 11 rocky planets mentioned in Section IIB. The unknown surface temperatures (or other parameters) can be estimated using various statistical models. In addition, parameters such as e.g. orbital period, stellar flux, distance of the planet from host star, etc. may be equally important to determine the habitability. Future work may include incorporating more input parameters to the Cobb-Douglas function, coupled with tweaking the attribute-enhanced K-NN algorithm by checking an additional condition. Cobb-Douglas, as proved, is a scalable model and doesn't violate curvature with additional predictor variables. However, it is pertinent to check for the dominant parameters that contribute more towards the habitability score. This can be accomplished by computing percentage contributions to the response variable – the habitability score. We would like to conclude by stressing on the efficacy of the method of using a few of the parameters rather than sweeping through a host of properties listed in the catalogs, effectively reducing the dimensionality of the problem.

To sum up, CD-HPF and CDHS turn out to be self-contained metrics for habitability. We would like pose the following questions in this context:

- How the two approaches coincide/converge?
- What is the implication in the overall scientific context?

The CDHS of Kepler 186-f planet turns out to be very close to the Earth's. The habitability potential of Kepler 186-f, estimated via Earth similarity, was computed by the explicit approach (Method 1, Figure 1). The implicit approach (Method 2, Figure 1) assigns Kepler 186-f the label of a habitable class of exoplanets. We observe the identical scenario in the case of the TRAPPIST-1 system, where the CDHS of the TRAPPIST-1 planets align with expected class labels.

Rapid discoveries of exoplanets notwithstanding, it is unrealistic and premature to predict how Earth-like are the conditions on any planet on the basis of the scant data. The best case scenario is to adduce a list of optimistic targets for future detailed missions. This manuscript achieves that goal by combining physical observations, mathematical rigor and machine learning techniques. However, our approach might pay rich dividends as encouraging observations have been reported very recently, [21]. Space Telescope Imaging Spectrograph (STIS) has been used to study the amount of ultraviolet radiation received by the TRAPPIST-1 planets. This helps determine the water content of the seven planets. The three planets within the star 's habitable zone, TRAPPIST- 1e, f and g may possess abundant amounts of water on their surfaces indicating habitability. We predict the same using machine learning and sophisticated modeling reported in the main paper and the supplementary file, [12]. By earth similarity approach (Method 1, Figure 1), we obtain the habitability scores of TRAPPIST-1e, f and g close enough to Earth (within two decimal places,Table VII). These planets are also classified as habitable by boosted tree learning (Method 2, Figure 1, Table VII). This is definitely encouraging.

References

[1] Bora, K., Saha, S., Agrawal, S., Safonova, M., Routh, S., Narasimhamurthy, A., 2016. CD-HPF: New Habitability Score Via Data Analytic Modeling. *Astronomy and Computing*, 17, 129-143

[2] Anglada-Escudé G. et al., 2016. A terrestrial planet candidate in a temperate orbit around Proxima Centauri. *Nature*, 536, 437-440.

[3] Witze, A. 2016. Earth-sized planet around nearby star is astronomy dream come true. *Nature*, 536, 381-382.

[4] Breakthrough Starshot. "A Russian billionaire has a crazy plan to reach a nearby planet that might harbor life". http://www.businessinsider.in, Retrieved 12 2016.

[5] Gillon, M., Triaud, A. H. M. J., Demory, B.-O., et al., 2017. Seven temperate terrestrial planets around the nearby ultracool dwarf star TRAPPIST-1. *Nature*, 542, 456-460.

[6] Schulze-Makuch, D., Méndez, A., Fairén, A. G., et al., 2011. A Two-Tiered Approach to Assessing the Habitability of Exoplanets. *Astrobiology*, 11, 1041.

[7] Rogers, L. A., 2014. Most 1.6 Earth-radius Planets are Not Rocky. *Ap. J.*, 801:1, 41.

[8] Saha S., Sarkar J., Dwivedi A., Dwivedi N., Anand M. N., Roy R., 2016. A novel revenue optimization model to address the operation and maintenance cost of a data center. *Journal of Cloud Computing*, 5:1, 1-23.

[9] Cobb, C. W. and Douglas, P. H., 1928. A Theory of Production. *American Economic Review*, 18 (Supplement), 139.

[10] Agrawal, R., Imielinski, T. and Swami A., Mining Association Rules between Sets of Items in Large Databases. In Proc. 1993 ACM SIGMOD International Conference on Management of Data, Washington DC (USA), pp. 207-216.

[11] Agrawal, R., Srikant, R.1994. Fast Algorithms for Mining Association Rules in Large Databases. In Proc. 20th International Conference on Very Large Data Bases (VLDB '94), Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo (Eds.). (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA), 487-499.

[12] Supplementary File: https://www.overleaf.com/read/gfkncnwxtsfm; Last Updated on 23/08/2017.

[13] Ginde, G., Saha, S., Mathur, A., Venkatagiri, S., Vadakkepat, S., Narasimhamurthy, A., B.S. Daya Sagar., 2016. ScientoBASE: A Framework and Model for Computing Scholastic Indicators of Non-Local Influence of Journals via Native Data Acquisition Algorithms. *J. Scientometrics*, 107:1, 1-51

[14] Chen T., Guestrin C., XGBoost: A Scalable Tree Boosting System, Preprint arXiv:1603.02754

[15] Zweig, M. H. and Campbell, G. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39 (4), pp. 561-577.

[16] Peres D. J., Iuppa C., Cavallaro L., Cancelliere A., Foti E., 2015, Significant wave height record extension by neural networks and reanalysis wind data, vol. 94, p. 128-140, doi:10.1016/j.ocemod.2015.08.002

[17] Van Rijsbergen C. J., 1979, Information Retrieval, ISBN:9780408709293

[18] http://www.aljazeera.com/indepth/opinion /2017/03/finding-earth-planets-important-trappist-1-nasa-170302100258232.html, Why discovering new Earth-like planets matters, as accessed on March 3, 2017.

[19] http://www.trappist.one/#news, as accessed on March 3, 2017.

[20] Safonova, M., Murthy, J., & Shchekinov, Y. A. 2016, Age aspects of habitability. *International Journal of Astrobiology*, 15, 93.

[21] V. Bourrier, J. de Wit, E. Bolmont, V. Stamenkovic, P.J. Wheatley, A.J Burgasser, L. Delrez, B.-O. Demory, D. Ehrenreich, M. Gillon, E. Jehin, J. Leconte, S.M. Lederer, N. Lewis, A.H.M.J. Triaud, V. Van Grootel, Temporal Evolution of the High-Energy Irradiation and water content of TRAPPIST-1 Exoplanets, 2017, Draft Version On Hubble.